C-CORE Publication No. 91-15

PERFORMANCE OF SELECTED SPEECH FEATURES FOR ISOLATED DIGIT RECOGNITION OF SPEECH BY A NEURAL NETWORK MODEL

Voon Hong Chin
C-CORE, Memorial University of Newfoundland
St. John's, Nfld., Canada. A1B 3X5

Abstract

An evaluation on the effectiveness of using the LPC, the PARCOR and the FFT coefficients as speech features input to the Learning Vector Quantization (LVQ) neural network model for digits speech recognition was performed. Speech samples from nine people were used for training and testing the network's performance. The result of the tests showed that the LPC and the PARCOR coefficients yielded better performance than the FFT coefficients.

Introduction

Though recent developments of sophisticated signal processing algorithms have made great strides in the difficult task of speech processing, many basic questions remain to be answered in the area of speech recognition. Problems associated with speaker independence, speaker rate, noisy environment, signal variability such as loudness and stress remain unsolved [1].

The emergence of neural network models which are designed to mimic some of the pattern matching capabilities of the human brain, and thus to adapt or "learn", are seen as a promising solution to the problems which plague conventional speech recognition techniques. Encouraging results were reported using neural networks for speech recognition [2-5].

This study implements the LVQ neural network model [4,5] to perform speaker-independent isolated digit speech recognition. The LVQ net is a supervised-training network. Its adaptive property allows it to be used in problems where little *a priori* knowledge of the pattern class distribution is available. The LVQ network is used in this study to perform adaptive speech vector quantization (VQ). VQ is a feature extraction technique whereby a codebook of multidimensional speech feature vectors is generated via a clustering procedure [1]. The codebook vectors are used as reference templates in the speech recognition process. The LVQ network is an optimal vector quantizer with an adaptive and local updating procedure of the codebook. This ability allows the codebook entries themselves to be updated if signal variations are detected [5].

The principal aim of the project was to investigate the performance of the LVQ net for isolated digit speech recognition using three speech features as input:

these are the FFT coefficients, the LPC coefficients [6], and the Partial Correlation (PARCOR) coefficients [6].

Methods of Features Extraction

For the FFT feature, the project uses the log magnitude $D(k)$ of the FFT coefficients as formulated in equation (1):

$$X(k) = (1/N) \sum_{n=0}^{N-1} x_n \, e^{-i2\pi nk/N}$$
$$D(k) = \log_{10} | X(k) | \qquad (1)$$

where $x(n)$ is the n^{th} time-series data sample, $X(k)$ is the k^{th} Fourier coefficient, N is the data length, and $\| \cdot \|$ denotes the magnitude function.

An efficient method of computing the LPC $a_M(p)$ and PARCOR $k(p)$ coefficients is available via Burg's algorithm and the Levinson's recursion [7]. The M^{th} order PARCOR coefficient k can be computed using equation (2):

$$k_M = \frac{-2 \sum_{n=M}^{N-1} e^f_{M-1}(n) \, e^b_{M-1}(n-1)}{\sum_{n=M}^{N-1} \left(\left| e^f_{M-1}(n) \right|^2 + \left| e^b_{M-1}(n-1) \right|^2 \right)} \qquad (2)$$

where $e_M{}^f$ is the forward residual error for order M and $e_{M-1}{}^b$ is the backward residual error for order M-1. The LPC coefficients are then calculated via Levinson's recursion:

$$a_M(p) = a_{M-1}(p) + k_M \, a_{M-1}(M-p)$$
$$a_M(M) = k_M \qquad (3)$$
$$| k_{M+1} | \leq 1$$

Experimentation

Figure 1 sketches the experimental setup used in this study. The setup can be divided into four main sections: data input, data acquisition, preprocessing, and neuro-processing by a neuro-computer board.

The speech waveforms were recorded and stored on a portable Pulser cassette

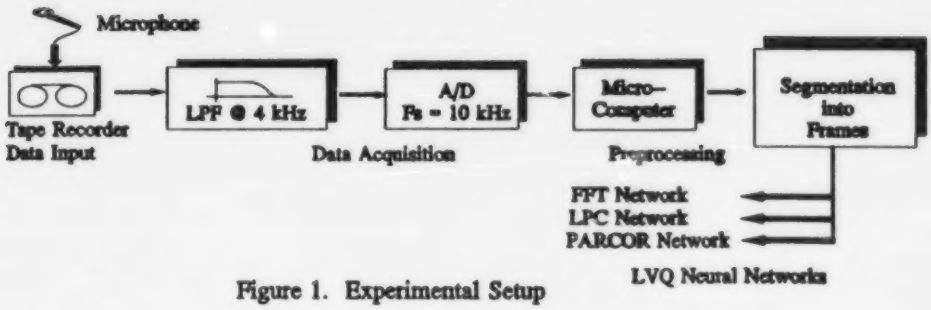

Figure 1. Experimental Setup

tape recorder via a condenser microphone. Both the microphone and recorder had a flat frequency response up to 20 kHz. The recorded waveforms were then fed into a 12-bit analog-to-digital (A/D) conversion board via a 4 kHz low pass filter to minimize aliasing. The A/D board sampled the waveform at 10 kHz.

In the preprocessing, each speech utterance was segmented into 15 time frames of 256 ms each. The d.c. bias on each time frame was removed before it was fed into the feature extraction algorithms. The log-magnitude of the FFT, the LPC, and the PARCOR coefficients were calculated using equations (1), (3), and (2). Only 126 frequency coefficients were used for the FFT input net, the d.c. and the last frequency coefficients were ignored. A 12^{th} order AR process was assumed for this study, giving 12 coefficients per time frame for the LPC and PARCOR input net.

A specialized neuro-computer board was used to implement an LVQ network for each input feature. The nets were called the FFT-net, the LPC-net, and the PARCOR-net. The FFT-net had 15 frames x 126 input nodes, 5 processing elements (PE) per category and 10 categories (for digits "0" to "9"). This gave a total of 1890 input nodes and 50 Kohonen PEs. Both the LPC and PARCOR input LVQ nets had the same network configuration: 15 frames x 12 input node (= 180 nodes), 5 PEs per category, and 10 categories.

For training the networks, a total of 70 data sets from 7 individuals were used. Each data set consisted of an utterance, e.g., "0" or "9". A training iteration using the 70 data sets through the network constituted an epoch. A total of 60 test samples were prepared from 6 individuals. The test data were taken from four individuals; composed of two individuals whose speech samples were used in training the net and two individuals whose speech samples were completely new to the net. The networks were tested after every 10 epochs of training.

Results and Conclusion

The results of the testing at various stages of the networks' training are given in Figures 2 and 3. Figures 2 (a) to (c) display the distribution of the number of

correct classifications with respect to the categories after 700 epochs. The test contained 6 samples for each category. Comparing Figures 2(a), (b) and (c), it can be seen that the LPC-net and the PARCOR-net both achieved greater than 50% correct classification scores. The FFT-net however performed poorly compared to the other two. This conclusion can also be supported by the result shown in Figure 3. Here, the overall test score (for 60 test samples) at various stages of the nets' training are plotted. These scores represent an average from all ten categories. The FFT-net's performance is consistently lower than that of the LPC-net and the PARCOR-net. The latter two features produced similar performances.

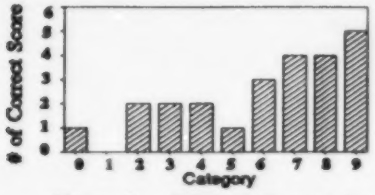

Figure 2a. FFT-net Correct Score
(6 Samples/Category: After 700 Epochs)

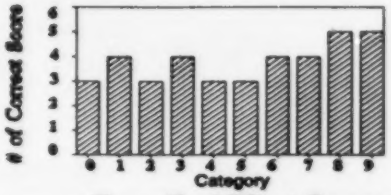

Figure 2b. LPC-net Correct Score
(6 Samples/Category: After 700 Epochs)

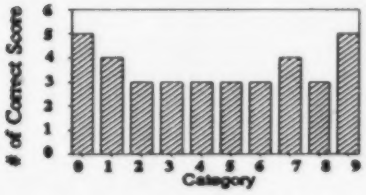

Figure 2c. PARCOR-net Correct Score
(6 Samples/Category: After 700 Epochs)

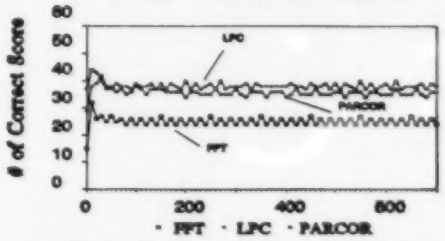

Figure 3. Overall Score for 60 Test Samples

Several factors could account for the poor performance of the FFT-net. One reason might be due to the large number of input nodes required which increased the dimensionality of the input space and hence the level of difficulties in performing the classifications. Another factor could be attributed to the larger ranges in the input data compared to those of the LPC and the PARCOR features. The large range generally causes difficulty in arriving at good solutions since dominant parameters (larger in coefficient values) associated with the speaker and not the speech, for instance, suppress or mask smaller variations found between different digit utterances.

From Figure 3, it is noted that the three nets converged after less than 50 epochs. This implies that the 700 epochs of training time would be sufficient to assess the effectiveness of the three speech features for the LVQ net as further iterations would yield the same conclusions. The results suggest that both the LPC and PARCOR coefficients as input to the LVQ network for digits speech recognition yield better performances than the FFT coefficients.

It should be noted that no attempt was made to improve the network's performance by fine-tuning it. This would include proper preprocessing of the input feature data such as scaling, and the determination of the optimal network parameters such as the number of PEs per category and the different learning rates. Also, a larger training sample size would improve the network's ability to learn and generalize better. These considerations were not included since they would not affect the performance evaluation objective of the study.

Acknowledgements

The author wishes to thank Dr. William Vetter for encouraging him to submit this paper and Dr. Jacques Yves Guigné for supporting the work. The editorial assistance of Eleanor Nesbitt-Friis is gratefully acknowledged.

References

[1] Levinson, S.E., and Roe, D.B., 1990. *A Perspective on Speech Recognition*, IEEE Communications Magazine, pp. 28-34, January.

[2] Bottou, L., Soulié, F.F., Blanchet, P., and Liénard, J.S., 1990. *Speaker-Independent Isolated Digit Recognition: Multilayer Perceptrons vs. Dynamic Time Warping*, Neural Networks, Vol. 3, pp. 453-465.

[3] Lang, K.J., and Waibel, A.H., 1990. *A Time-Delay Neural Network Architecture for Isolated Word Recognition*, Neural Networks, Vol. 3, pp. 23-43.

[4] Martinelli, G., Ricotti, L.P., and Ragazzini, S., 1990. *Nonstationary Lattice Quantization by a Self-Organizing Neural Network*, Vol. 3, pp. 385-393.

[5] Kohonen, T., 1988. *An Introduction to Neural Computing*, Neural Networks, Vol. 1, No. 1, pp. 3-16.

[6] Saito, S., and Nakata, K., 1985. *Fundamentals of Speech Signal Processing*, Academic Press, New York, pp. 94-125.

[7] Roberts, R. A., Mullis, C. T., 1987. *Digital Signal Processing*, Addison-Wesley publishing Company, Ontario, pp. 506 - 532.